The influence of Twitter data on the viewership of Banijay's show

Dominik Szewczyk 224180

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

Breda University OF APPLIED SCIENCES

The influence of Twitter data on the viewership of Banijay's show

Dominik Szewczyk
Breda University of Applied Sciences
Data Science and Artificial Intelligence
Nitin Bhushan (Lecturer)
20 January 2023 17:00

Introduction to Banijay company

Banijay company is home to over 120 production companies across 22 territories, and a multi-genre catalogue boasting over 146,000 hours of original standout programming.

The main specialization of Banijay is to create various types of shows popular around the world.
Company thanks to partnering to grow ideas and maximise on them worldwide, they have nurtured a growing catalogue spanning 146,000 hours of premium, high-quality content
Banijay is proud to be able to partner with any buyer or seller worldwide and subsequently provide talent with a home where they can be creatively free.
They are a lean team with big ambitions. Company encourages autonomy, promote independent thinking and decision-making, while driving a sense of unity through a shared collaborative spirit. And while there is one goal, they believe in the power of enabling and trusting their teams to use different paths to get there.
Banijay company try their best to create the most compelling content. They believe in taking risks to succeed and by fostering our regional knowledge to build innovative universal super brands and premium scripted hits, they have become a go-to for quality, no matter the language, no matter the genre and no matter the platform.
One of the company's statements:
“Banijay is a home for everyone, and together, we are committed to ensuring we have a truly representative and inclusive workforce that thrives on diverse perspectives. We want everyone to feel welcome and excited to work with us both on and off-screen.
Our planet is incredibly important to us and as such, we are taking the necessary steps to reduce our carbon footprint and our overall impact on the environment. Both corporately and onset, we are making changes to transform our business, all while also educating our audiences.”

Data analysis and machine learning are more and more often used in large companies. This allows you to adjust the show to the viewer and improve the company's results.
I analysed the data available to me to improve the viewership of the show and to determine who, when and why watches the programs. In the future, this should help producers better assess what should be on the show and when it should air. The analysis that allows us to determine what social group watches the program will help to adapt show better to viewers.

Nowadays, social media has an increasing influence on our lives. However, is it possible to predict the number of views based on the number of likes on Twitter?

Datasets used

In my project and analysis, I used 3 sets of data that I combined, cleaned, and understood.

1. Content dataset,
2. Ratings dataset,
3. Twitter dataset.

1. Content dataset is the data we received from Banijay. They contain information about show hosts, titles, ids and keywords, information about show starts and end time allowed us to combine this data with ratings data.
2. Rating data provides more information about the viewership of the show, divides the viewers into different **Target Group** (a target group is a subgroup of the population based on certain characteristics, for example age, gender and/or social class. The primary group of people that something, usually an advertising campaign, is designed to appeal to. In the data you will see for example “boodschapper _25_54”. That is the group of people who say that they must take care of all the tasks in their household (cooking food, grocery shopping, cleaning) between the ages of 25 and 54 years. “Boodschapper 25-54” is translated to shopper 25-54 in English.). In dataset, we can also find information about **Frequency** (whether the program is a rerun or not. In the data provided you will see only “live/prerecorded uitzendingen” which are the first run episodes.) and the most important **Rating** (the average number of people in a target group who watched a program. People who see or hear a particular program). Most of my charts and analyzes are based on this database.
3. The latest dataset was taken from Twitter. We used Twitter's Developer Platform for this. The platform provides tools, resources, data, and API products for us to integrate, and expand Twitter's impact through research, solutions and more. This API allows us to find and retrieve, engage with, or create a variety of different resources including the following: Tweets. Users. Spaces.

Data cleaning, and preparation and exploration

The large amounts of data we have is not everything, first the data must be cleaned, prepared, and understood

Starting with content data, I got to know the information about the data, such as the number of columns and rows, the number of empty rows, the number of duplicates, and finally I understood the data in general. I went to data cleaning, removed empty values and duplicates, translated the necessary columns from Dutch to English, changed the format to data-time in some columns, split the id column into show_id and fragment which allowed for better data visualization.

	id		show_id	fragment
0	OP1_____-WON02197428_01_segment	0	OP1_____-WON02197428	1_segment
1	OP1_____-WON02197428_02_segment	1	OP1_____-WON02197428	2_segment
2	OP1_____-WON02197428_03_segment	2	OP1_____-WON02197428	3_segment
3	OP1_____-WON02197428_04_segment	3	OP1_____-WON02197428	4_segment
4	OP1_____-WON02290378_01_segment	4	OP1_____-WON02290378	1_segment
...	...	...	...	...
2979	OP1_____-WON02251309_03_segment	2979	OP1_____-WON02251309	3_segment
2980	OP1_____-WON02251309_04_segment	2980	OP1_____-WON02251309	4_segment
2981	OP1_____-WON02340053_01_segment	2981	OP1_____-WON02340053	1_segment
2982	OP1_____-WON02340053_02_segment	2982	OP1_____-WON02340053	2_segment
2983	OP1_____-WON02340053_03_segment	2983	OP1_____-WON02340053	3_segment

These changes allowed me to combine content data and ratings data, I did it based on the start and end time of the show that was in content data and the time that was in ratings data

Target Audience Analysis

What is the average number of people who see or hear a particular program?
Thanks to data analysis, we know that the average viewership of the show is 208 viewers.

What is the average number of people per target group who see or hear a particular program?
Comparing the average viewership in particular groups allow us to conclude that people over 50 watch the show significantly more than other age groups

Target Group	
13_19_jr	5.587279
20_34_jr	22.673923
35_49_jr	58.784992
50plus_jr	538.793347
6_12_jr	3.391958
boodschapper_20_49	51.736065
boodschapper_25_54	70.626938
m_6plus_jr	305.638966
tot6plus	626.427932
v_6plus_jr	320.788525

Content Analysis

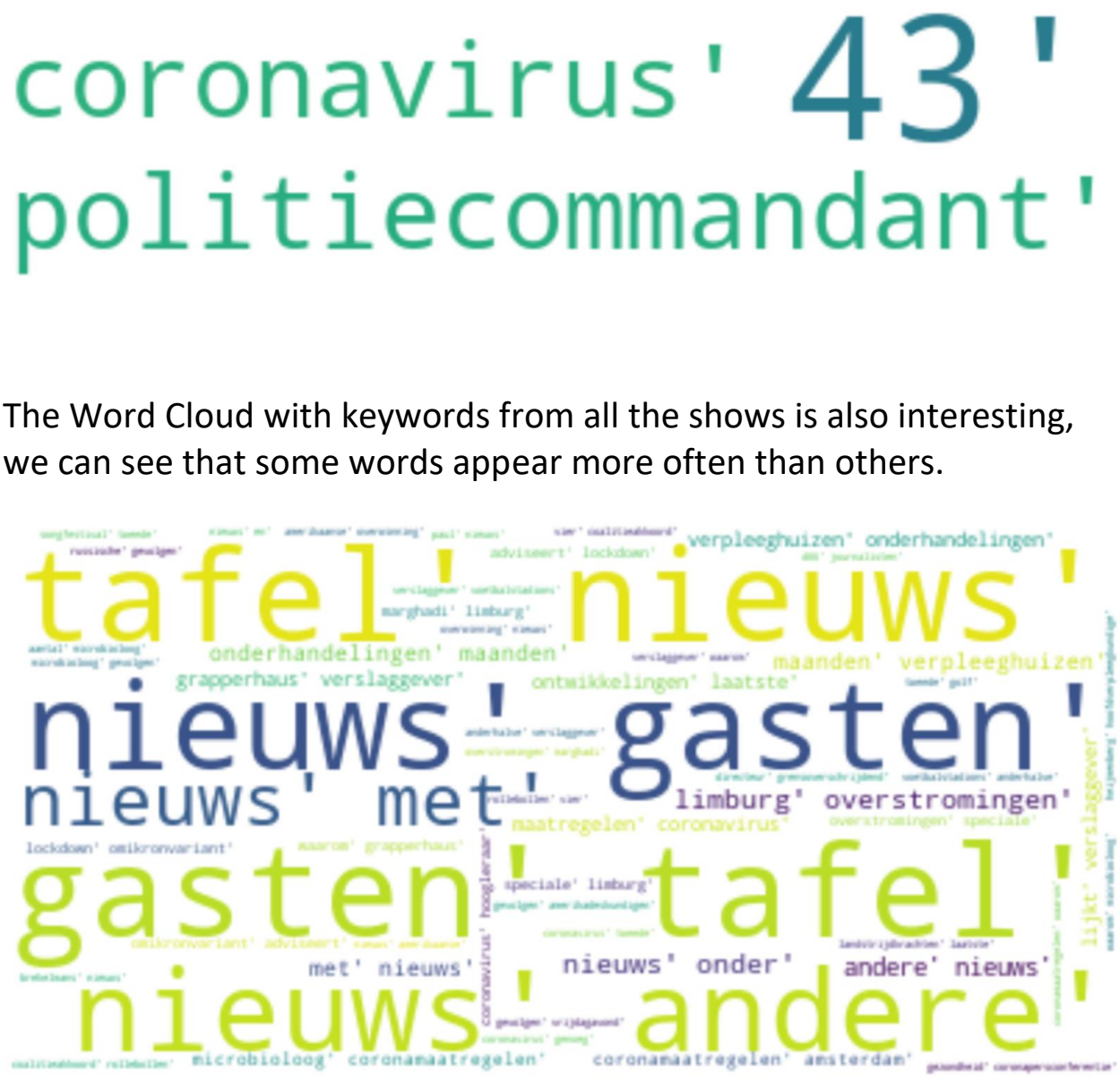
Who are the most highly rated hosts?
Below we see the 5 top rated hosts:

['Pauw, Jeroen', 'Ekiz, Fidan']	326.537392
['Napel, Carrie ten', 'Groenhuijsen, Charles', 'Sijtsma, Welmoed']	322.920066
['Ekiz, Fidan', 'Pauw, Jeroen']	318.015574
['Ostiana, Giovanca', 'Brink, Tijs van den', 'Fikse, Margje']	311.296935
['Sijtsma, Welmoed', 'Groenhuijsen, Charles']	302.774795

What are the most highly rated shows and what are the most highly rated fragments in those shows?
Below we see the 5 top rated shows with the most highly rated fragments:

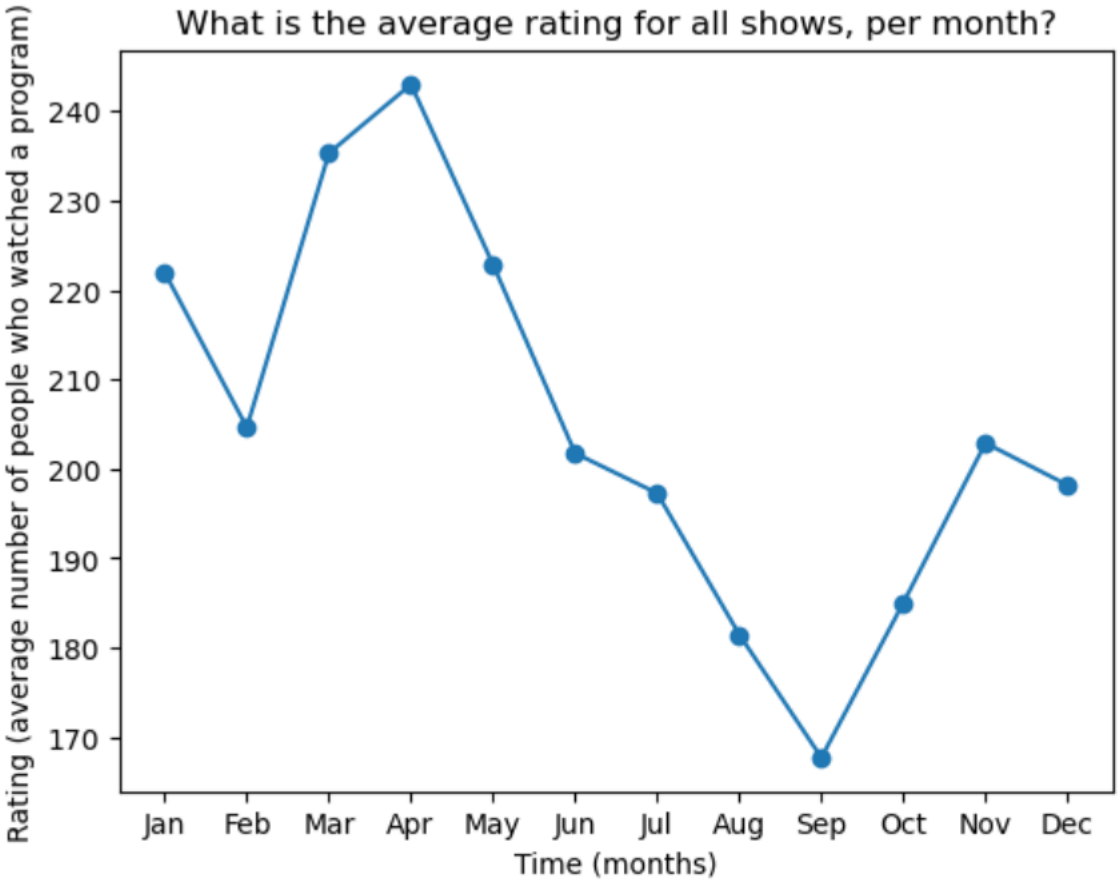
Ab Osterhaus, Rob Jetten en Max Daniel over of de politiek meer in actie moet komen	1
488.208214	
Het LUMC werkt aan virusremmer tegen het coronavirus	2
466.870394	
Hoe waarschuw je kinderen en jongeren over het coronavirus? Joris Marseille en Saskia Smith vertellen het	3
459.579559	
Laila Frank, Raymond Mens, Maarten van Rossem, Tijs van den Brink en Kirsten Verdel over de huidige situatie in Amerika	1
443.206767	
Eppo van Nispen met de mooiste historische sneeuwbeelden	3

The question is what caught the attention of viewers? Thanks to the data and keyword analysis, I created a word cloud with the most popular keywords of these shows.

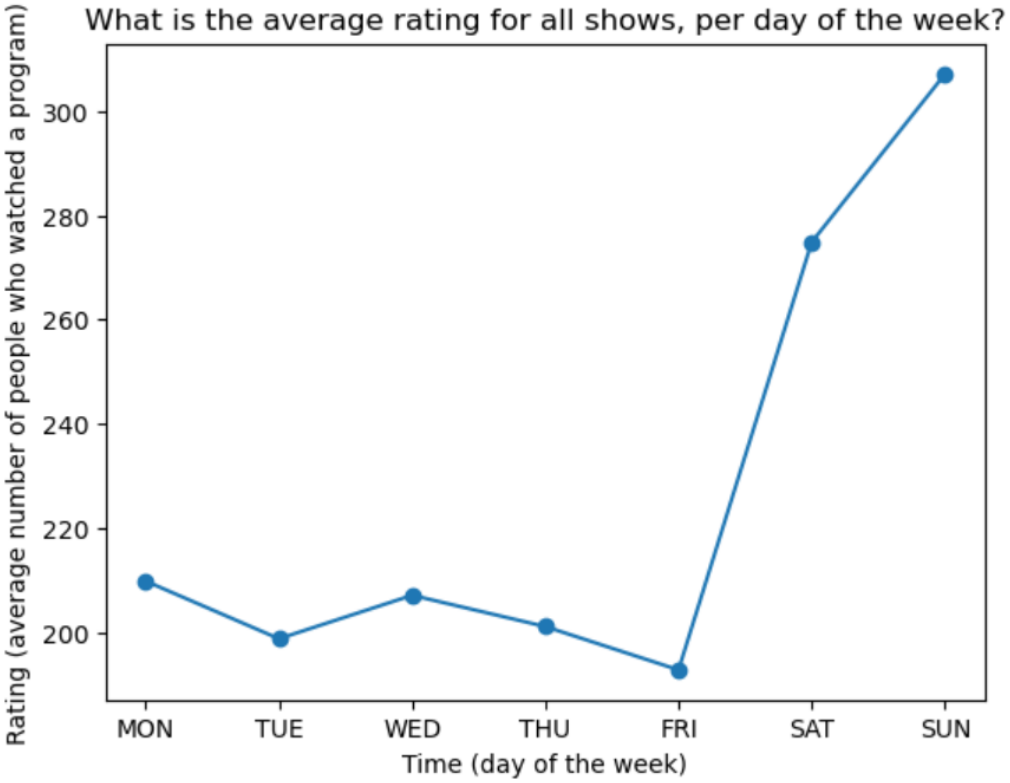


The Word Cloud with keywords from all the shows is also interesting, we can see that some words appear more often than others.

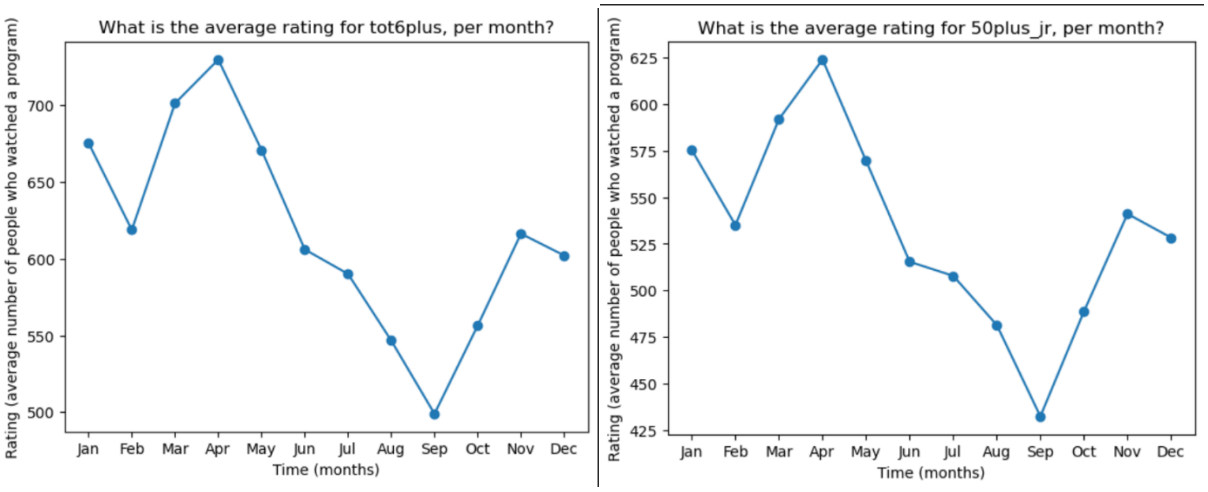
Trend Analysis



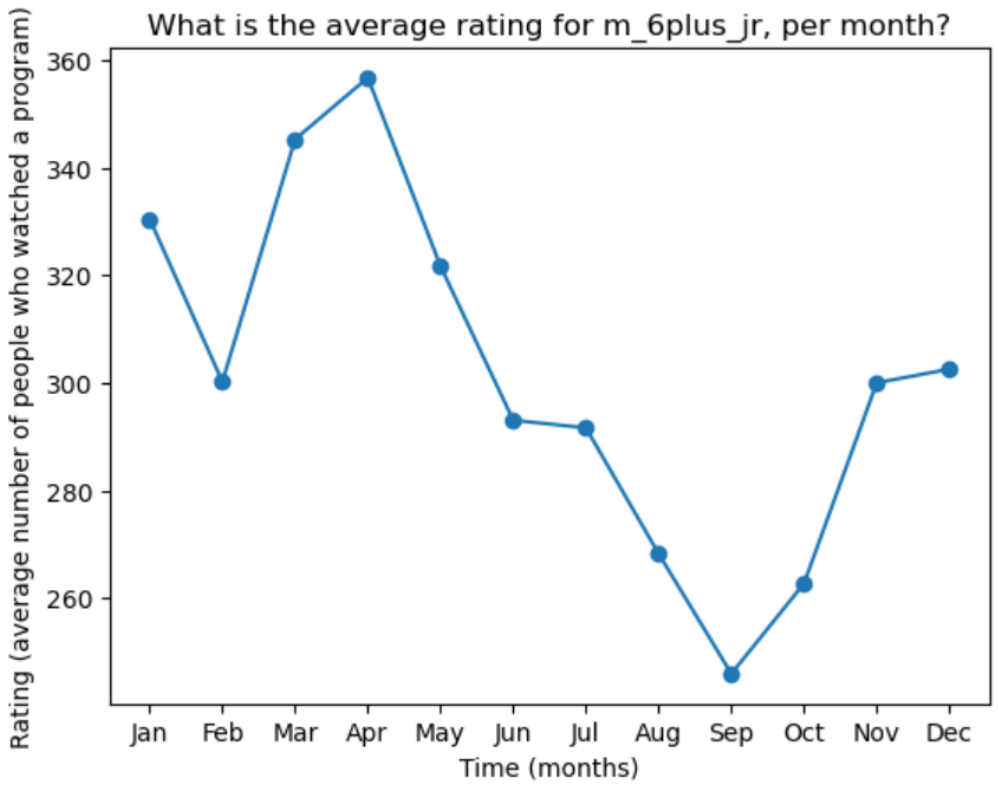
In this chart we can see what the viewership was each month. At the beginning of the year, in January, the average viewership was 220 viewers, in February it dropped to 205 and then increased until April (over 240 viewers). Then we can see a gradual decline in viewership until September, when the viewership was the lowest with only 170 viewers. By the end of the year, viewership increased again to 200 viewers



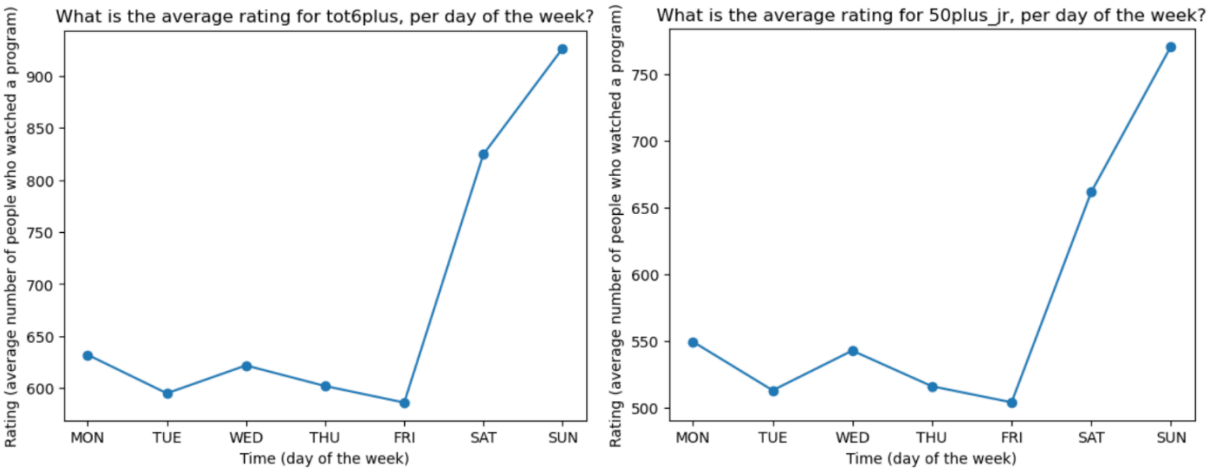
The chart above shows us what the viewership was depending on the day of the week. We can see that there is a huge difference between the weekend and other days of the week. On Saturday, the show is watched by 280 viewers, and on Sunday this number increases to over 300, which is almost 30% more than on other days. This means that the best shows should be broadcast on weekends because that is when most people will watch them.



During analysis for target groups with the highest viewership, we can notice a similarity. The number of viewers is the highest in April and then decreases significantly until September. During these 5 months, the average viewer drops by as much as 200 from 625 viewers to 425 viewers.



In this target group (m 6 plus jr) we can see that in October viewership is lower than in other target groups. We can see that the increase from September to October is lower than the average increase in these months in other target groups



We can see a relationship between the average viewership and the average viewership in the top target groups, and here and there the viewership increases sharply on Saturday and decreases on Monday.

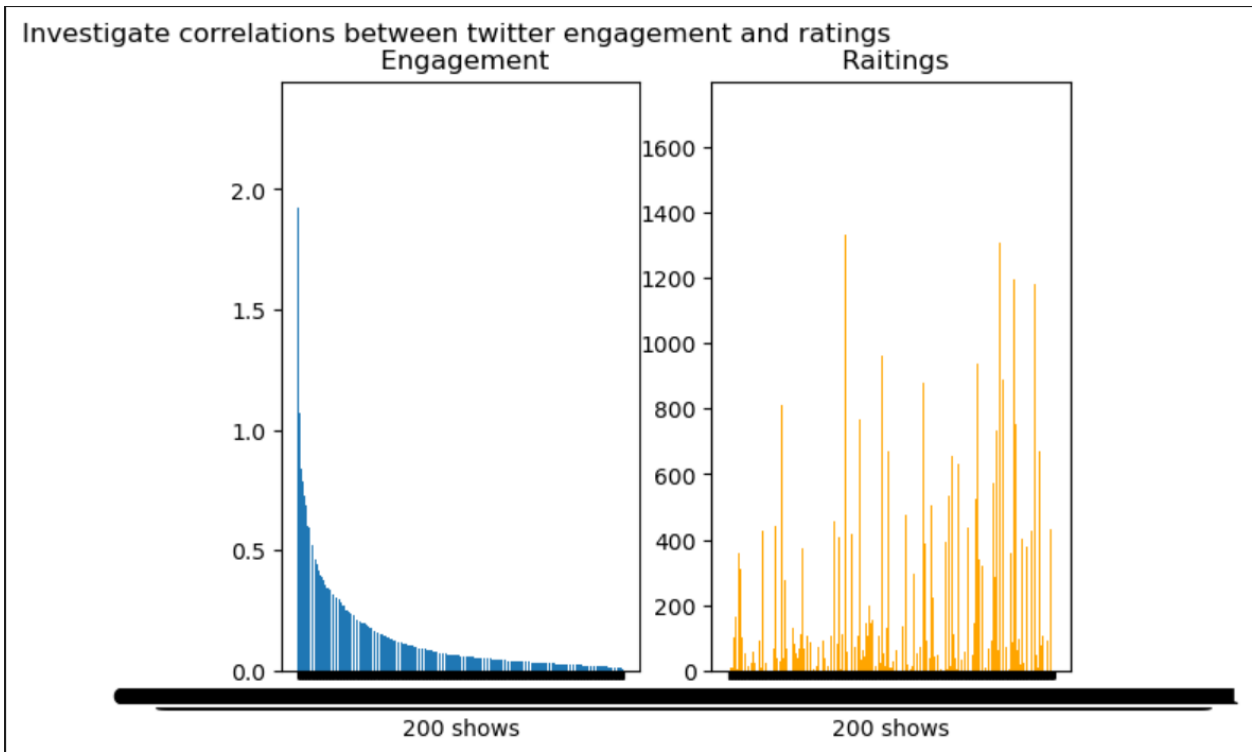
Social Media Analysis

Nowadays, social media is very important in business. Social media can help us to engage with our customers and find out what people are saying about our business. They are also a source of information on the interests of recipients and prevailing trends, this information can be used to satisfy the viewer.
I started analysing the Twitter data by understanding the columns and rows, then I changed the format of the columns where it was necessary, next I only kept those tweets that do not reference another tweet, thanks to this, I retained only the most necessary information.
With all the necessary data, I calculated the engagement rate, which are the most important thing in marketing nowadays because it indicates the popularity of the content. Engagement rate can help you discern which posts are performing well, what content is resonating with your audience, and how to effectively communicate your success, no matter how many followers your account has. Engagement rate provides a way to determine growth without vanity metrics, such as follower count. To calculate it, you need to add up the number of retweets likes replies and total quite tweets and divide it by number of followers.

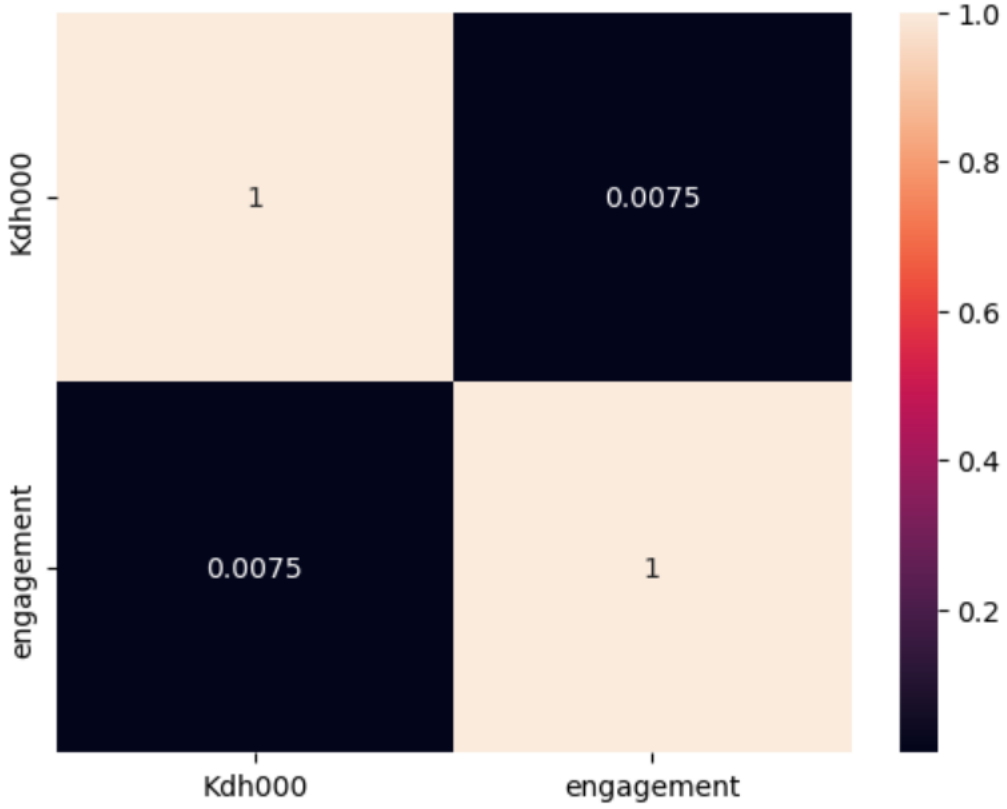
Below we can see the 5 shows with the highest engagement rate

id	engagement	title
OP1_____-WON02323453_01	2.328838	Op1 - De opening van maandag 13 juni
OP1_____-WON02275165_01	1.922404	Op1 - De opening van woensdag 1 december
OP1_____-WON02145145_01	1.669347	Politiek verslaggever Marloes Lemsom en Dieder...
OP1_____-WON02306111_01	1.315423	Op1 - De opening van 5 april
OP1_____-WON02113058_01	1.069782	Kinderarts-immunoloog Emmeline Buddingh over v...

Engagement rate should depend on viewership, so I decided to check it out. I combined all the datasets I had and compared the viewership column and the engagement rate column. Below we see a visualization of the results for the first 200 shows.



You can clearly see that there is no correlation, but to be sure, I calculated the correlation coefficient and visualized it as well. (Kdh000 = column with ratings)



Machine learning model

With the use of machine learning (ML), which is a form of artificial intelligence (AI), software programs can predict outcomes more accurately without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Supervised learning is a subcategory of machine learning and artificial intelligence. It is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be separated into two types of problems when data mining—classification and regression:

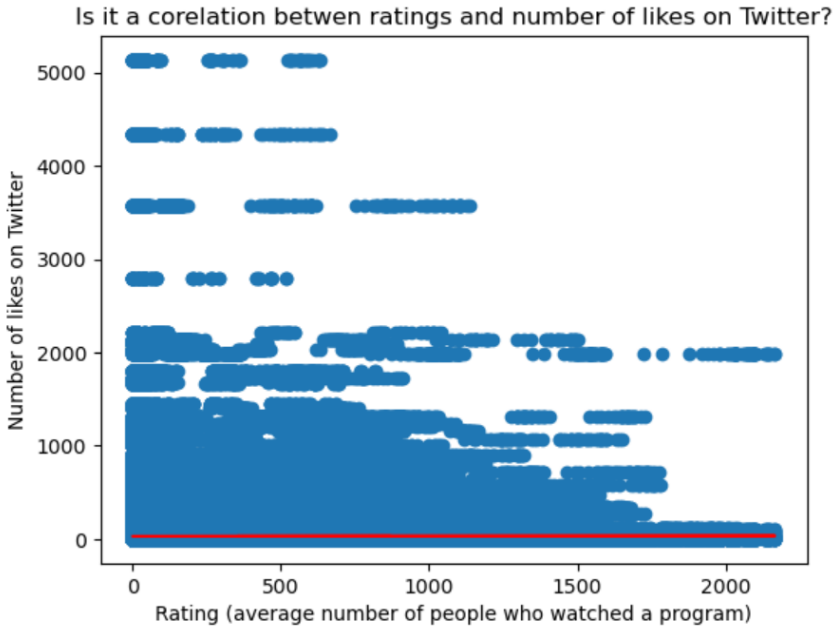
Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labelled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbour, and random forest, which are described in more detail below.
Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

One of the models that performed better used linear regression. Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes. However, unlike other regression models, this line is straight when plotted on a graph.

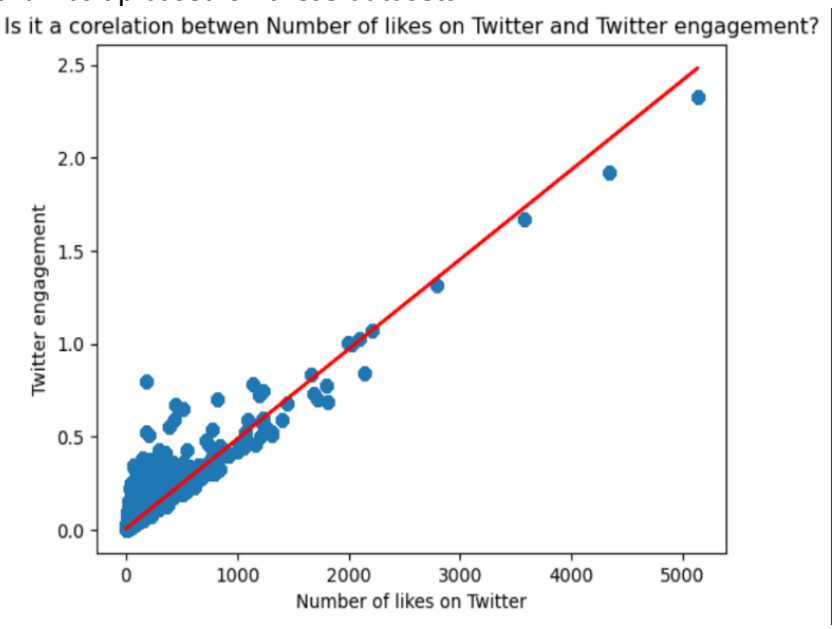
I tested 2 machine learning models: decision tree and linear regression, unfortunately, none of these models gave me the accuracy that would allow me to base conclusions on it

In my project, I used the regression method to create a program that allows you to predict the number of viewers based on the data of the target group or the number of likes on Twitter. It would help to create a show that more people watch.

Trying to predict the number of views based on the target group, I got the accuracy of the model at the level of 54%, which means that the prediction would not be accurate and cannot be used. that's why I tried with a statement of the number of views and the number of likes on Twitter. I decided to visualize the results and add a linear regression to see the trend. Unfortunately, also in this case the results were not satisfactory.



Further analysis allowed me to notice that the number of likes on Twitter is related to Twitter engagement and as I have previously proved, there is no correlation between viewership and Twitter engagement, so it is impossible to predict the number of viewers based on Twitter's thumbs up based on these datasets



Ethical part

Considering the ethical side of the project, I remembered that its goal was to increase human intelligence, I wanted the data that I would analyse to allow the Banijay company to increase its viewership, I also remembered that the data must belong to the creator, I did not share the data anywhere and used it only for the needs of the project and very it is important that the IA model is understandable and transparent, in the project I showed how my model was created step by step and I tried to describe it in such a way that it was transparent.

Together with the Banijay company, we made sure that all data was collected legally, respecting the rules of the GDPR. I also read the privacy policy of twitter to ensure that all my actions are ethical.

To improve the ethics of the project, you could check if the age given by the viewers is correct and is appropriate to use the twitter from which we download data for our project.

Results and conclusion

While analysing the data, I noticed that the most frequent viewers are people from the target group tot6plus and 50plus_jr, shows produced for them should gain the highest viewership.
The most popular hosts are: [“Pauw, Jeroen,” “Ekiz, Fidan”] and [“Napel, Carrie ten,” Groenhuijsen, Charles,” “Sijtsma, Welmoed”]
Shows containing keywords such as ‘coronavirus’ ‘43’ and ‘politiecommandant’ are viewed the most.
Another relationship is that people most often watch the show in April and least often in September. In fact, based on the data, it also turned out that the highest viewership is on weekends
Analysing the data from Twitter, it turned out that Twitter engagement has no correlation with the number of views. Machine learning models have confirmed this, and we probably need more precise data for further research

Is it possible to predict the number of views based on likes on twitter?
No, because there is no relationship between the number of likes and the number of views, the reason for this is probably low activity on twitter, and developing a profile on twitter would help to collect more data, thanks to which it would be possible to predict viewership. I think the company should pay attention to this and try to do something about it.

References

1. *Home*. (2022, December 7). Banijay Group - We Are Banijay. https://www.banijay.com/

2. *What is Supervised Learning? | IBM*. (n.d.). https://www.ibm.com/topics/supervised-learning

3. Access *Denied*. (n.d.). https://business.adobe.com/blog/basics/business-case

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility

Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

PHONE
+31 76 533 22 03
E-MAIL
communications@buas.nl
WEBSITE
www.BUas.nl

DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES